

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problem Mailbox.**

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平5-274471

(43)公開日 平成5年(1993)10月22日

(51)Int.Cl. ⁵ G 0 6 K 9/20 G 0 6 F 15/62 H 0 4 N 1/38	識別記号 3 4 0 J 序内整理番号 3 3 0 G 8125-5L 4226-5C	F I	技術表示箇所
---	---	-----	--------

審査請求 未請求 請求項の数1(全8頁)

(21)出願番号 特願平4-71903	(71)出願人 000004226 日本電信電話株式会社 東京都千代田区内幸町一丁目1番6号
(22)出願日 平成4年(1992)3月30日	(72)発明者 水町 肇 東京都千代田区内幸町1丁目1番6号 日本電信電話株式会社内

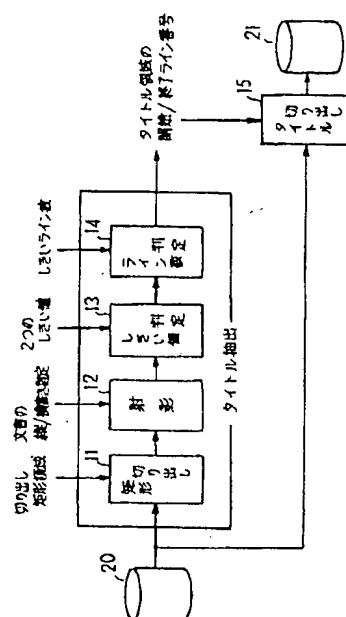
(74)代理人 弁理士 森田 寛

(54)【発明の名称】 イメージ文書のタイトル領域抽出処理方法

(57)【要約】

【目的】 本発明は、入力されたイメージ文書データから、文章のタイトル領域を自動的に抽出する手段を提供することを目的としている。

【構成】 イメージ文書上の指定された矩形領域について、イメージ・データについての例えば水平方向への射影をとり、当該射影値が所定のしきい値範囲内にある横行が所定個数連続する所を、タイトル領域として抽出する。



1

【特許請求の範囲】

【請求項1】 イメージ文書中のタイトル領域を自動的に抽出する処理方法において、
入力されたイメージ文書データから指定された矩形領域を切り出す領域切り出し手段と、
前記の領域切り出し手段より切り出された矩形領域について、当該矩形領域の縦方向もしくは横方向についての射影を、指定に応じて算出する射影算出手段と、
前記の射影算出手段により算出された各ライン毎の射影値が、指定された2つのしきい値の範囲内か否かを判定するしきい値判定手段と、
前記のしきい値判定手段により判定されたラインの連続ライン数が、指定されたしきいライン数以上か否かを判定するライン数判定手段とを有し、
イメージ文書データから文書タイトルが書かれている領域を抽出するようにしたことを特徴とするイメージ文書のタイトル領域抽出処理方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、ファクシミリ装置やイメージ・スキャナ装置から、白黒2値のイメージ・データとして入力されたイメージ文書データを蓄積し、蓄積されたデータを対話的に検索する装置において、検索に当たりイメージ文書の内容を的確に表現する文書中のタイトル領域を自動的に抽出し、検索時に複数のタイトル領域を表示し、検索を容易にするようにしたイメージ文書のタイトル領域抽出処理方法に関するものである。

【0002】

【従来の技術】 従来、イメージ文書の検索に当たって、文書を識別するためのタイトル情報として、ファクシミリ装置から入力する場合には送信相手や受信時刻またはマークシートにより付加された情報が、またイメージ・入力装置から入力する場合にはマークシートや別途キーボードなどから入力された付加情報が夫々利用されている。

【0003】

【発明が解決しようとする課題】 しかし、送信相手や受信時刻を用いた場合には文書の内容が判断できず、また、マークシートやキーボードを用いる場合には付加的な情報を付与するための作業が必要となるなどの欠点があつた。

【0004】 本発明は、従来のタイトル情報の付与方法の上記欠点を改善するため、入力されたイメージ文書データから、タイトル強調のために書かれた罫線や、文書の右上に書かれた日付け等を誤認識することなく、文章のタイトル領域を自動的に抽出することを目的としている。

【0005】

【課題を解決するための手段】 入力されたイメージ文書データから指定された矩形領域を切り出す領域切り出し

2

手段により文書の右上に書かれた日付けや左上に書かれた宛先などをタイトル領域の抽出のための処理範囲外とすることで誤抽出を防止し、イメージ文書が縦書きか横書きかに応じて前述の領域切り出し手段より切り出された矩形領域について、当該矩形領域の縦方向もしくは横方向の射影を算出する射影算出手段により算出された各ライン毎の射影値が、指定された2つのしきい値の範囲内か否かを判定するようにし、当該しきい値判定手段よりノイズを除去すると共に射影値が2つのしきい値の範囲内にあるラインを判定することで背景が白の上に黒でタイトル文字が書かれた場合と背景が黒の上に白でタイトル文字の何れの場合でもタイトル文字を検知可能にする。

【0006】

【作用】 前述のしきい値判定手段により判定されたラインの連続ライン数が指定されたしきいライン数以上か否かを判定するライン数判定手段によりタイトル文字の強調のために書かれた罫線などを誤抽出を防ぎ、最初にライン数判定手段によりしきいライン以上と判定された連続ライン域をタイトル領域として抽出するようする。

【0007】 従来の技術とは、マークシートやキーボード入力によりタイトル情報を付加することなく、入力されたイメージ文書データから自動的にタイトル領域を抽出することが異なる。

【0008】

【実施例】 以下、図面を用いて本発明の実施例を説明する。図1は本発明の一実施例に係るイメージ文書のタイトル領域抽出方法の構成を示すブロック図であり図2、図3は図1の各ブロックでの処理過程を示した処理態様を示す。

【0009】 図中の符号11は矩形領域切り出し部、12は射影部、13はしきい値判定部、14はライン数判定部、15はタイトル切り出し部、20はイメージ文書蓄積装置、21はタイトル蓄積装置を表している。

【0010】 また22はイメージ文書、23は指定された矩形領域(切り出し領域)、24は切り出された矩形領域、25は算出された射影値、26は判定結果、27はタイトル領域を表す。

【0011】 入力されたイメージ文書が蓄積されているイメージ文書蓄積装置20から処理対象とするイメージ文書22と指定された矩形領域23とを矩形領域切り出し部11への入力とし、矩形領域切り出し部11により切り出された矩形領域24のイメージ・データを出力する。

【0012】 この出力されたイメージ・データと文章の縦／横書き指定とは射影部12の入力となり、射影部12は文章の縦／横書き指定に基づき縦方向もしくは横方向の各ライン毎のヒストグラム値を算出する。該当ラインの全ての画素がビット・オフの場合の値を0.0、全ての画素がビット・オンの場合の値を1.0として算出され

3

たヒストグラム値を0.0から1.0までの値の範囲で正規化することにより射影値25を算出する。

【0013】この射影値25と指定された上限値および下限値の2値のしきい値とはしきい値判定部13の入力となり、しきい値判定部13は各ラインの射影値が上限値および下限値の2値のしきい値の範囲内であれば該当ラインの判定フラグをオンに、また範囲外であれば判定フラグをオフに設定した判定結果26を出力する。

【0014】この判定結果26としきいライン数とはライン数判定部14の入力となり、ライン数判定部14-10は、文章の文頭から文末までの方向にライン毎の判定フラグが連続するライン数を求め、最初にしきいライン数を越える領域を判定することでタイトル領域27を検出し、その開始ライン番号と終了ライン番号とを出力する。

【0015】この開始ライン番号と終了ライン番号に基づき、タイトル切り出し部15は、対象のイメージ文書データの先頭からタイトル領域の開始ラインまで読み飛ばし、開始ラインから終了ラインまでのイメージ・データをタイトル領域として切り出し、タイトル蓄積装置220-1に出力してタイトル情報として蓄積する。

【0016】尚、イメージ文書蓄積装置20およびタイトル蓄積装置21は、磁気ディスク装置などの二次記憶媒体を利用している。図4は図1に対応する処理フローを示す。そして図5は図4におけるステップ1(STEP1)に対応するフローチャート、図6は図4におけるステップ2(STEP2)に対応するフローチャート、図7は図4におけるステップ3(STEP3)に対応するフローチャート、図8は図4におけるステップ4(STEP4)に対応するフローチャートを示す。

【0017】図4において、

(ステップ1)：図1に示す矩形領域切り出し部11の処理に対応しており、図2に示す切り出された矩形領域24を得る。

【0018】(ステップ2)：図1に示す射影部12の処理に対応しており、図3に示す射影値25を得る。

(ステップ3)：図1に示すしきい値判定部13の処理に対応しており、図3に示す上限しきい値HHと下限しきい値HLとの範囲に存在するものを抽出する。

【0019】(ステップ4)：図1に示すライン数判定部14の処理に対応しており、図3に示す判定結果26にもとづいて、図3に示すタイトル領域27を得る。

(ステップ5)：図1に示すタイトル切り出し部15の処理に対応している。

【0020】以下、図5ないし図8を参照しつつ具体的に説明する。

処理(1)：図2に示す切り出し領域23における左上端の座標P₁(x₁, y₁)と右下端の座標P₂(x₂, y₂)とを取得する。

【0021】

4

処理(2)：P₁とP₂との座標値の差を計算する。

$$d_x = x_2 - x_1$$

$$d_y = y_2 - y_1$$

処理(3)：イメージ文書の元データの各ドットが例えば多値データで表される場合などにおいては、イメージ文書の元データが複数枚(複数プレーン)で与えられることとなり、最初の処理に当たっては第1枚目であることを指示すべく i_p = 1 にセットする。

【0022】処理(4)：i_p = i のソース(原文-イメージ文書) S R C における x 座標が 1 から x₂ までと y 座標が y₁ から y₂ までの範囲のイメージを、 ; = i 用のデスクネーション(作業用の保持領域) D E S における x 座標が 1 から x₂ までと y 座標が 1 から y₂ までの範囲の所に転記する。

【0023】処理(5)：i_p = i について処理(4)が終ると i_p = (i + 1) のソースについての処理を行うべく i_p = i_p + 1 にする。

処理(6)：i_p が上記プレーンの枚数と一致するか否かを比較する。i_p が大であればターンに進み ; が大でなければ処理(4)に戻り、 ; = (i + 1) のソースについての転記を行う。

【0024】処理(7)：射影値の算出を行うべく行(横行)に対応するカウント値 i を i = 1 にセットする。

処理(8)：射影値 SUM(i) をゼロにクリアする。

【0025】処理(9)：第 i 行の桁に対応するカウント値 j を j = 1 にセットする。

処理(10)：プレーンの第1枚目から処理すべく i_p = 1 にセットする。

処理(11)：射影値 SUM(i) について上記作業用の保持領域上のあるドットの値 D E S(p, j, i) を加算する。即ち、

$$SUM(i) = SUM(i) + DES(p, j, i)$$

を演算する。

【0026】処理(12)：次のプレーンの処理に移るべく i_p = i_p + 1 を行う。

処理(13)：i_p とプレーンの枚数とを比較する。 ; が大であれば処理(14)に進み、 ; が小であれば、次のプレーンの処理に進むように処理(11)に戻る。

【0027】処理(14)：次の桁位置の処理に入るべく j = j + 1 を行う。

処理(15)：j と d_x とを比較する。 j が大であれば処理(16)に進み、 j が小であれば、次の行についての処理に進むべく処理(10)に戻る。

【0028】処理(16)：次の行の処理に入るべく i = i + 1 を行う。

処理(17)：i と d_y とを比較する。 i が大であれば全プレーンについての全行について算出されたこととなりリターンに進み、 i が小であれば、次の行についての射影値の算出処理に入るために、処理(8)に戻る。

5

【0029】処理(18)：上限しきい値HHと下限しきい値HLとを与える。

処理(19)：将来において正規化を行うための準備として、プレーン枚数とcとの積をd_{max}として得ておく。これは、処理対象となり得る全行の全桁がすべて論理「1」であったとしたときの値_{max}を得るものである。

【0030】処理(20)：最初の行についての処理を行なうべくi=1にセットする。

処理(21)：第i行についての射影値SUM(i)をd_{max}で割って、正規化された値hを得る。

【0031】処理(22)：HHがhと比較される。HHが小であれば処理(25)に進み、HHが大であれば処理(23)に進む。

処理(23)：HLがhと比較される。HLが大であれば処理(25)に進み、HLが小であれば処理(24)に進む。

【0032】処理(24)：第i行の正規化値hがHHとHLとの間にあり、第i行についての判定フラグHFLAG(i)を「1」にする。

処理(25)：第i行についての判定フラグHFLAG(i)を「0」にする。

【0033】処理(26)：次の行についての処理に移るべくi=i+1にする。

処理(27)：iとd_yとを比較する。iが大であればリターンに進みiが小であれば、次の行についての処理に進むように処理(21)に戻る。

【0034】処理(28)：しきいライン値Lを与える。

処理(29)：第1行目から処理に入るべくi=1にセットする。

処理(30)：カウント値countをゼロにリセットする。

【0035】処理(31)：上記第i行の判定値HFLAG(i)が「1」か「0」か調べられる。

処理(32)：「1」の場合、カウント値countをプラス1する。

【0036】処理(33)：「0」の場合、カウント値countを「0」にする。

処理(34)：次の行についての処理に進むためにi=i+1を行う。

6

処理(35)：Lがcountと比較される。Lが小であれば処理(36)に進み、Lが大であれば次の行についての処理に移るために処理(31)に戻る。

【0037】処理(36)：切り出し位置をy₁から(count+y₁)までとする。

ここで、タイトル抽出を行う対象を横書き印刷されたイメージ文書などのように限定すれば、外部から指定する、切り出し矩形領域、文書の縦／横書き指定、2つのしきい値、1つのしきいライン数を固定値とすることができる。

【0038】

【発明の効果】以上説明したように、本発明によれば、入力されたイメージ文書から自動的にタイトル領域が抽出可能となるため、イメージ文書を蓄積するための入力操作の省力化、および、蓄積されたイメージ文書の検索において表示された複数のタイトルから必要な文書を選択させることによるイメージ文書の検索操作の効率化が図れる。

【図面の簡単な説明】

【図1】本発明の一実施例に係るイメージ文書のタイトル領域抽出方法の構成を示すブロック図である。

【図2】図1の各ブロックでの処理過程を示した説明図である。

【図3】図1の各ブロックでの処理過程を示した説明図である。

【図4】全体のフローチャートを示す

【図5】個々のフローチャートを示す。

【図6】個々のフローチャートを示す。

【図7】個々のフローチャートを示す。

【図8】個々のフローチャートを示す。

【符号の説明】

1 1 矩形領域切り出し部

1 2 射影部

1 3 しきい値判定部

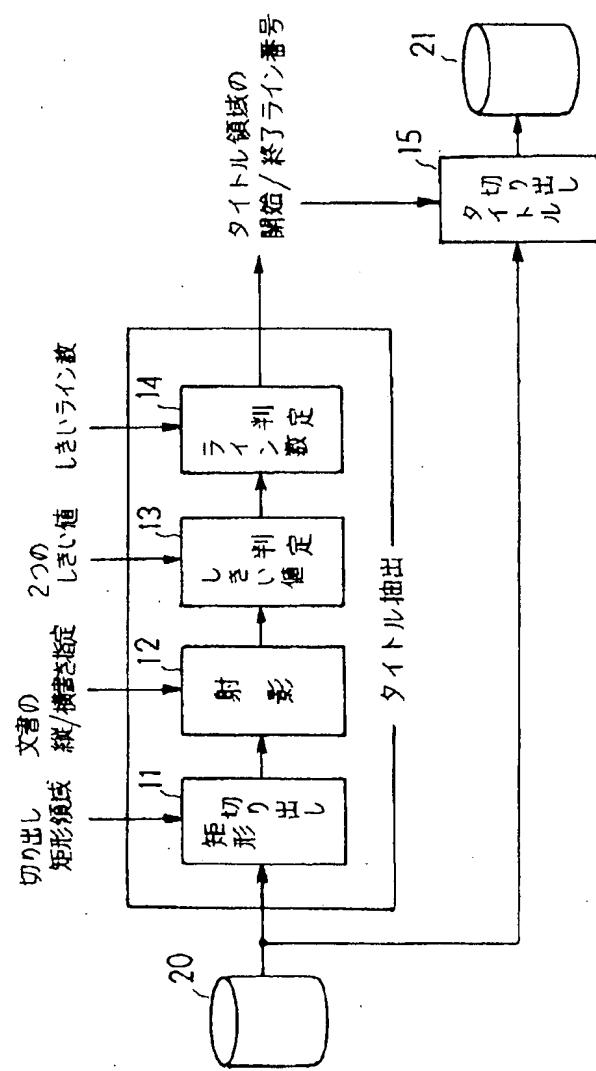
1 4 ライン数判定部

1 5 タイトル切り出し部

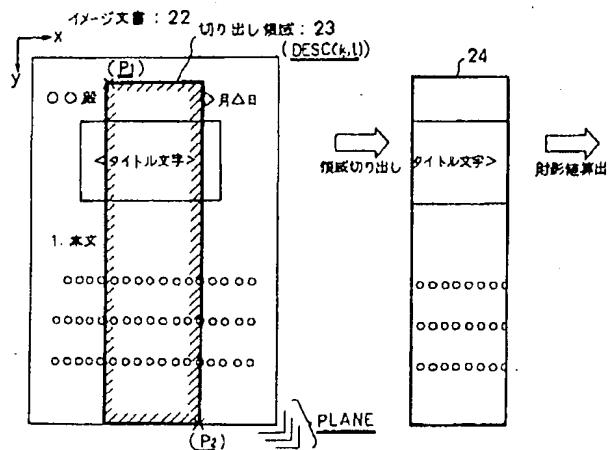
2 0 イメージ文書蓄積装置

2 1 タイトル蓄積装置

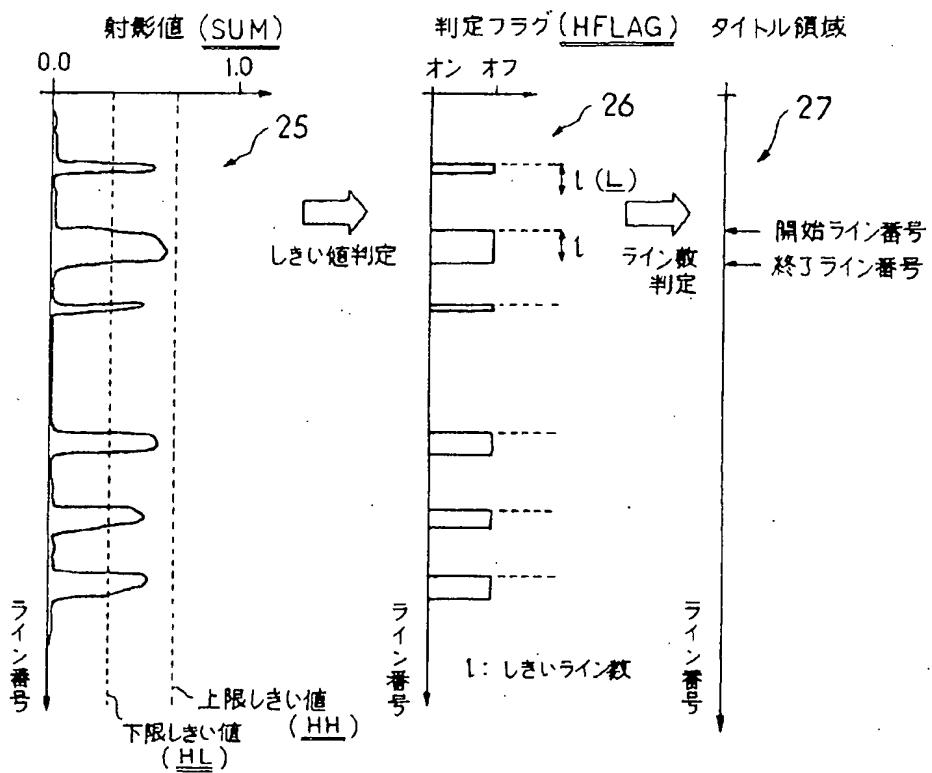
【図1】



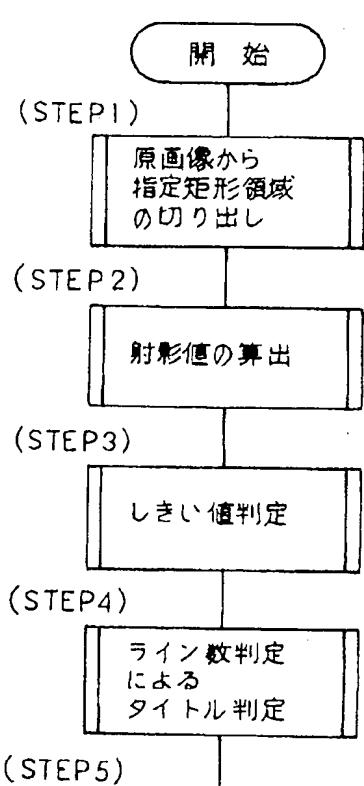
【図2】



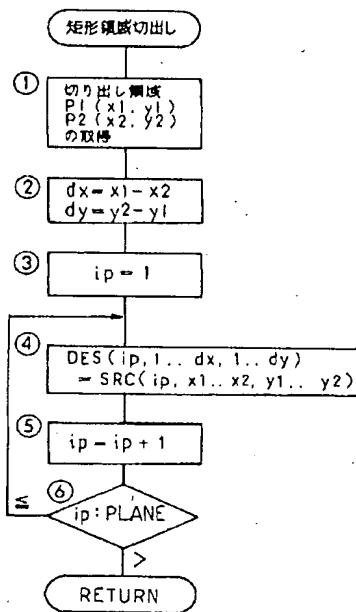
【図3】



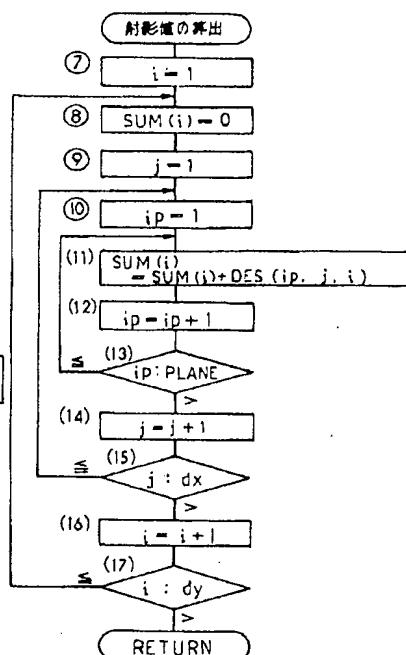
【図4】



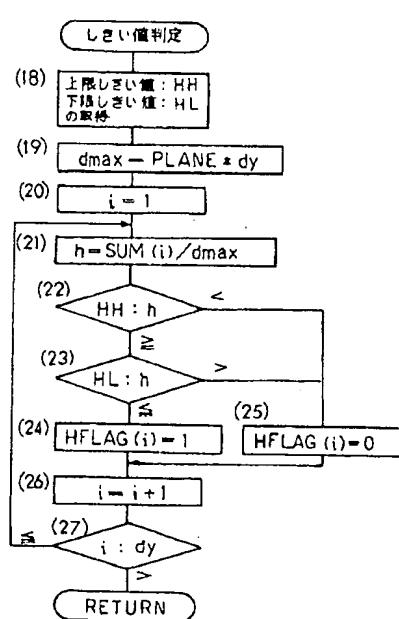
【図5】



【図6】



【図7】



【図8】

